What evidence should guidelines take note of?

Nancy Cartwright FBA, FAcSS, Professor of Philosophy

Philosophy Dept, Durham Univ, Durham DH13HN, UK

Running title: what evidence

Correspondence to: above address, ph (44) 01865 761688, fax 44 (0)191 334 6551

Key words: causal inference, methods, RCTs, guidelines, causal Bayes nets, case studies, process tracing, instrumental variables, causal modelling

Abstract

The Guidelines Challenge Conference on which this special issue builds asked as the first of its 'further relevant questions': '*How do we incorporate more types of causally relevant information in guidelines?*' This paper first supports the presupposition of this question – that we need further kinds of evidence – by pointing out that the RCT, touted as the best source of evidence on effectiveness, can do so little for us. Second, it outlines a number of other good ways to learn what will work that the medical community, and much of the public health community, are not making much use of.

Introduction

The Guidelines Challenge Conference on which this special issue builds asked as the first of its 'further relevant questions': '*How do we incorporate more types of causally relevant information in guidelines?*' I am not going to offer suggestions on *how* to incorporate more types though. Rather I shall first support the presupposition of this question – that we need further kinds of evidence – by pointing out that the RCT, touted as the best source of evidence on effectiveness, can do so little for us. Second, I shall outline a number of other good ways to learn what will work that the medical community, and much of the public health community, are not making much use of.

When it comes to tools for predicting effectiveness the usual suspects in evidence-based medicine (EBM) are randomized controlled trials (RCTs). Central to the argument that RCTs can warrant causal conclusions and especially for the claim that they can estimate effect size (or average treatment effect – *ATE*) is the potential outcome equation (*POE*). A POE describes the causal possibilities that can affect a given outcome in an individual. Knowledge of the full POE for an individual or a population of individuals would be incredibly useful in predicting outcomes for them. Yet the RCT provides only a small amount of information about the POE, and that only for the study population, which is seldom the population of interest. Given these limitations, it is surely a good idea also to use other methods that allow us to draw causal conclusions, and often conclusions about the population of interest. I'll briefly describe a number of these in section 4.

1. *The POE*

A potential outcome equation for outcome Y for individual i loos like this:

$$\text{POE: } Y_i = \beta_i T_i + \sum_{j=1}^{J} \gamma_j x_{ij}$$

T_i is a dichotomous (1,0) treatment variable indicating whether *i* is treated and β_i is the *individual treatment effect* of the treatment on *i*: it represents how much T contributes to the outcome *Y* for individual *i*, which depends on the the factors that regulate this. The *x*'s are other linear causes of the outcome. The POE is causal through and through. It is supposed to capture a minimal set of causes of Y_i sufficient to fix its value for i. Philosophers and epidemiologists will probably be familiar with the same ideas in a Boolean yes/no-variable version:

$$\text{Boolean POE: } Y_i \equiv (T_i \,\&\, A_{i1}{}^1 \,\&\ldots\&\, A_{in}{}^1) \vee\ldots\vee (C_i{}^k \,\&\, A_{i1}{}^k \,\&\ldots\&\, A_{im}{}^k)$$

The POE assumes that causes are INUS conditions: *i*nsufficient but *n*ecessary parts of *u*nnecessary but *s*ufficient conditions for a contribution to the effect.

- Each disjunct/summand is sufficient but not necessary for the effect or for a contribution to the effect. This is neatly summarized in the slogan "There's more than one way to skin a cat."
- Within a cluster represented as a disjunct/summand, no factor by itself can produce an effect. All the elements are required to get a contribution to the effect. This reflects the idea that the salient cause – say, a proposed treatment – needs "help" to produce its contribution. These helping factors are called "moderator", "support" or "interactive" factors and get charted in familiar epidemiologists' pies.

2. *The trouble with RCTs*

The trouble with RCTs is that they don't tell us much. RCTs investigate only one aspect of the POE : β_i; then, only a population average of it: Exp β_i; and for that, they give only an estimate of it; this estimate has the virtue of being unbiased, which may not be so useful, and may lack precision, which normally is useful; and the estimate is for one specific population – the population enrolled in the trial. This last is not peculiar to RCTs of course: study results are always about the things that are studied, not about something else. Going beyond the study population requires other knowledge, generally much other knowledge.

So, let's look at this aspect to see what RCTs are good for. Ideally we would like to learn the individual treatment effect: the value of Y that individual i would have if treated minus the value i would have for Y if not treated "everything else constant": $Y_T(i) - Y_{-T}(i)$. By inspection, we can see that this difference is represented by β_i. Supposing *orthogonality*,[1] the observable difference in means between the outcome in the treatment group and in the outcome group is an *unbiased* estimate of the ATE:

$$\text{Exp}\,[Y_T(i) - Y_{-T}(i)] = \text{Exp}\,\beta_i$$

That's pretty amazing: we can estimate the average across a set of numbers without knowing even one of the numbers. Derivatively from this result we can draw explicit causal conclusions: if the ATE is positive, it follows that the treatment must cause the outcome in at least some individuals in the study population. What fixes the value of the ATE? From its position in the POE we see that β_i moderates the contribution T makes to Y, hence is a function of the net effect of the support factors for T to produce Y in i. So the ATE depends on the distribution of support factors in a population.

There are two main drawbacks to RCTs for predicting what will work. First, orthogonality is hard to achieve. Randomization is supposed to achieve this at base but much can go wrong after and most experiments are neither blinded at all points where it could matter nor well policed for correlations that arise post-randomization from other difference the two groups experience, like time, place, and length of treatment, clinician skill, etc.

Second are issues of bias versus precision. That the difference in average between treatment and control groups is an unbiased estimate of the ATE: i.e., they match in expectation over indefinitely many repeated randomizations on the study population. But generally we do only a single run on the study population.[2] What we would probably prefer is precision – getting close to the answer. We would get exactly the right answer if the net effect of all causal factors other than the treatment were exactly balanced between the two wings of the experiment. It is common in the evidence-

[1] T is probabilistically independent of β,Y,x.

[2] Moreover, given that individuals change over time, this is the most one could do.

based medicine literature to defend RCTs on the grounds that they compare like with like. That's what would happen if the two wings of the experiment *were* balanced. In that case the observed difference in means would be the actual ATE, not just an estimate of it.

So precision is about balance and RCTs do not balance anything. Indeed, balance in a trial is improbable. Of course with bigger samples we get better balance. But what happens, for instance, if there is only one important cause, it is unknown, unobservable, and unbalanced? Consider a case described in Deaton and Cartwright [1] where treatment effects over individuals have mean zero and are distributed as a shifted lognormal distribution – there are asymmetric treatment effects, as perhaps with expenditures on healthcare where most people have zero expenditure in any period, but a few individuals spend huge amounts that account for a large share of the total. Although the true ATE is zero, in Monte Carlo runs of the experiment significant effects occur too often, and the problem persists at quite large sample sizes, though it improves. These results are driven by the outliers: the estimate of the ATE depends on whether the outlier is in treatment or control.

What, then, in the end do we learn? We get an *unbiased estimate* of the answer, *not the answer*. But anyway, what's the question? We estimate the ATE in the study population. This is a function of the average of the net effect of the support factors – as they are distributed in the study population, which is very likely to vary across populations. So there's no way to avoid the need to learn what the support factors are. We need to know their actual values for individual patients to make predictions about outcomes for them, and we need to know their distribution in any new population to predict the average effect there. This is well known. But then it is startling how much more emphasis is put on the RCT than on studies that help identify these other important facts without which the RCT result is of no use. Any conclusions beyond the results for the population enrolled in the study depend *entirely* on other knowledge, and as Deaton & Cartwright [2,12] note: 'A rigorously established result whose use elsewhere is justified by a loose declaration of simile is no stronger than a number pulled out of the air.'

None of this is improved by doing a number of RCTs and agglomerating results in some way or another. That could be helpful if the study populations were drawn randomly from the target or if we could assume the distribution of support factors is similar over all the populations, which we seldom have good reasons to suppose. Otherwise we are just getting information about the mean of the distribution of support factors in more and more populations, none of which are the target.

3. When, and how, do RCT results travel?

The ATE in a population depends on the average net effect of the support factors for the treatment in that population. Hence it should be the same in two populations when the support factors have the same distribution. To transport ATE as is, you need warrant for this assumptions. Alternatively you can reweight. Take the ATEs conditional on the various different values of β and multiply them by the probabilities of those values in the new population to get the ATE in the new population -- suppopsing you have warrant for all those values you assume.

There is more besides this that can be done, as work by Judea Pearl and Elias Bareinboim shows [3,4]. Pearl and Bareinboim suppose we have available both causal information and probabilistic information for the experimental population; for the target we have only probabilistic information. We also suppose that certain probabilistic and causal facts are shared between the two and certain ones are not. Pearl and Bareinboim derive theorems describing what causal conclusions about the target population are fixed given these facts. This shows which similarities and difference between

two populations allow which experimental results from one to be used in which ways to calculate different probabilistic and causal facts in the other. So, your warrant for transporting a result from one population to another, as is or adjusted in various ways is just as strong as your warrant that these background assumptions are all true. Securing this is a tall order, requiring a huge mix of methods. Guidelines that ignore this will give bad advice. The exception are cases where it can be assumed that everything that matters is the same in the necessary ways. But there is generally no justification for taking that as a default position.

Unfortunately the demands outstretch even what we learn from Pearl and Bareinboim. Their framework supposes that the relations between effect and causes can be represented in potential outcomes equations. If the POEs for the two populations are different with respect to the role of the treatment T – with respect to the very capacity of T to affect Y – transportability fails. As they say, in cases where 'the target domain does not share any mechanism with its counterpart…. the only way to achieve transportability is to identify R [the causal relation of interest] from scratch in the target population' [3, 588].

What capacity a treatment has to contribute to an effect for an individual depends on the underlying structures – physiological, material, psychological, cultural and economic – that makes some causal pathways possible for that individual and some not, some likely and some unlikely. This is a well-recognised problem when it comes to making inferences from model organisms to people. But it is equally a problem in making inferences from one person to another or from one population to another. Yet in these latter cases it is too often downplayed.

When the problem is explicitly noted, it is often addressed by treating the underlying structures as moderators in the potential outcomes equation: give a name to a structure-type – men/women, old/young, poor/well off, from a particular ethnic background, member of a particular religious or cultural group, urban/rural, etc. Then introduce a yes-no moderator variable for it. Formally this can be done, and sometimes it works well enough. But giving a name to a structure type does nothing towards telling us what the details of the structure are that matter nor how to identify them. In particular, the usual methods for hunting moderator variables, like subgroup analysis, are of little help in uncovering what the aspects of a structure are that afford the causal pathways of interest. Getting a grip on what structures support similar causal pathways is central to using results from one place as evidence about another, and a casual treatment of them is likely to lead to mistaken inferences. The methodology for how to go about this is under developed, or at best under articulated, in EBM, possibly because it cannot be well done with familiar statistical methods and the ways we use to do it are not manualizable. It may be that medicine has fewer worries here than do social science and social policy, due to the relative stability of biological structures and disease processes. But this is no excuse for undefended presumptions about structural similarity.

4. Taking a direct approach

Rather than a roundabout approach, trying to learn about one population by studying another, we could try studying the target population itself. The usual wisdom is that the best thing to do is to conduct an RCT there. That would be so if the RCT can be done on a representative sample of the population, the relevant characteristics of the population do not change between the time of the RCT and the time that matters for what we want to do, all we need is an unbiased estimate of the ATE, the underlying distribution of individual treatment effects has a nice shape so our estimate will be reasonably precise (and we have warrant for assuming this), randomization is easy to do, we have

enough knowledge of the other causal factors to guard against post-randomization correlations, and the RCT is morally permissible and not more expensive than it is worth. If any of these conditions fail, there's plenty one can do instead. Here I will briefly describe some of the options.

All methods require assumptions to justify the results they are supposed to show. And all methods for drawing causal conclusions require causal assumptions ("No causes in, no causes out"). Some of the methods I describe here are *clinchers*. For these it is *provable* that, if the related assumptions met, the conclusion follows. For RCTs, the central assumptions are that the causal possibilities in the population studied can be represented in a potential outcomes equation, that the study treatment is orthogonal to all causes of the outcome other than its own downstream effects, and (to estimate precision) assumptions about the shape of the underlying distribution of individual treatment effects. It follows deductively from these that the difference in means for the outcome between the treatment and the control groups is an unbiased estimate of the population ATE. Different clinchers need different inputs and deliver different outputs. Other methods can provide evidence that can help make a case for a conclusion even if the conclusion doesn't follow deductively. I call these *vouchers.*

The distinction between clinchers and vouchers assumes that what a method consists in can be reasonably well articulated. The description of the method lays out the kinds of conclusions the method can address and the assumptions that have to be met if it is to do so. Clinchers secure the conclusion – if the assumptions are met; in principle all the uncertainty rests in whether the assumptions are met. A voucher, by contrast, only *speaks for* the conclusion. Even if it is done to the letter and all the requisite assumption are true, a voucher does not establish its conclusion. The familiar case in EBM are statistical methods that establish a correlation. These provides *evidence* for a causal conclusion, but this evidence must be combined with evidence of different kinds from other methods to secure the conclusion.

There is no way to draw a firm distinction between clinchers and vouchers. It all depends on how we articulate the methods. Consulting a fortune teller can be a clincher – so long as the method instructs us to consult a fortune teller who gets right answers. Still it is a useful distinction since good methodology dictates articulating methods in such a way that we know how to apply them, which means that we should have a good idea what it takes to warrant the method's assumptions.

4a. Clinchers

Instrumental variables provide an unbiased estimate of the ATE in a population using observational data from the population by identifying "instruments" that affect the treatment variable but have no effect on the outcome other than via the treatment:

Instrumental variable(s) → Treatment variable(s) → Outcome measure

Distance from hospital → Use of treatment A or B → Recovery from surgery

The instrument is like an experimental intervention that changes the cause under test and no other causes of the outcome, so that any changes in the outcome can be attributed to the putative cause. These have been advocated for use in medicine and public health by, among other, JP Newhouse & M McClellan [5]. Julian Reiss [6] provides an excellent discussion of the method and of the conditions that must be met for its conclusion to be secured.

Causal structural models are a set of POEs, one for each of a set of time-ordered variables, which will have a triangular (or block triangular) form:

x_1
$x_2 = a_{21}x_1$
$x_3 = a_{31}x_1 + a_{32}x_2$
...

Standard econometric techniques can, in happy circumstances, estimate the coefficients in these equations from observational data on a population. In that case the equations should be functionally correct for the population. That does not make them causal–real POEs where only causes of the dependent variable appear on the right-hand side. They could just represent associations, "correlations". There are, though, special conditions in which, provably, a functionally correct structural model is causally correct. These conditions are related to those that must be met for the equations to contain an instrumental variable. It may be rare for these conditions to be satisfied, but when they are we can get a great deal of causal information about a population from observational data. For more on this, see Cartwright [7, also in 8].

Causal Bayes Nets methods, developed by Pearl and associates at UCLA and Clark Glymour and associates at Carnegie Mellon, derive new causal information about a population from available causal and probabilistic information ("correlations") from that population. They suppose that causal relations can be represented in POEs with a probability measure over the variables caused outside the system, and that they can be graphed in related directed acyclic graphs (DAGs), with causes at the heads of arrows and effects at the feet. Supposing that the causal relations satisfy some widely applicable assumptions (that are mostly shared with RCT design), these methods can produce every DAG that is consistent with the input information. One drawback is that DAGs do not distinguish disjuncts from conjuncts in the INUS formulae, so they do not picture which factors figure as part of the same sufficient cause complex and which act independently or in other complexes, which QCA aims to do. For a general description of Bayes nets methods see Spirtes [9]; for an illustration of how to use them to reduce bias due to confounding in a medical context see Shrier and Platt [10]; for reservations about their usefulness in those medical contexts where causation is more effectively represented with differential equations than with POEs, see Aalen, et. al. [11].[3]

Deduction from theory is widely used to draw causal conclusions throughout engineering and the natural sciences. Clearly the warrant for the conclusion depends on the warrant for the premises, but for the premises as they need to be formulated to bear on the case at hand. A generally well-warranted theory that has not been much tested in similar cases might provide less warrant than one that's been very successful in predictions of similar conclusions in similar settings even if the theory does not have such a good track record elsewhere. I am often told that this approach is not useful for Guideline construction because we have so few trustworthy theories to guide us in medicine and public health. This is a relative matter. How trustworthy are the premises required for the theory deduction of the conclusion in view as compared to the premises required for this or that other method to deliver that same conclusion? I would suppose that it is not often that the premises needed for any methods are highly trustworthy. If so, there's no way to avoid having to figure out what lessons to draw from a mix of evidence all of which is dicey.

[3] I am not sure how widespread this problem is. I have not, for instance, seen a defence of the RCT as a method for causal inference or in cases where causality is represented in a dynamic equations framework.

4b. Vouchers

Case studies are commonly used in legal cases, ethnographies, policy evaluation and post-hoc fault diagnoses. They employ a mix of qualitative and quantitative methods and can often credibly establish cause-effect relations for individuals but are seldom useful for estimating ATES in a population since generally the right kinds of information will not be available for every individual to allow a credible judgment one way or the other about causation.

These may be of special help in studying rare diseases, where statistics are not available. But their usefulness is not confined to rare diseases since they document trajectories, which can be very useful for clinical care in many other diseases. One of the major movements in health care recently is the use of patient records to document variations in trajectories and outcomes using data mining of massive patient record data bases, particularly in the US given the way insurance documents all episodes, tests, interventions etc., but also in the European Union's drive to fund research into future and emerging technologies - health informatics. Diabetes, where we are concerned about management by patient and clinician through a life course, is a key illustration. For one among many examples of a diabetes case study see Preuveneers & Berbers [12]. For a good general account of case study methods, see Byrne and Ragin [13].

Qualitative comparative analysis (QCA) scours population data to provide the full functional form for the Boolean version of a POE-like equation for an outcome of interest, thus attempting to identify both moderator variables and the causes that act independently of the treatment. It is a voucher because without a great deal of causal input there is no way to see whether the resulting equation is a genuine POE – truly causal – or a mere statement of association ("correlational"). As with instrumental variables and causal structural modelling, additional assumptions could be built riht into the description of the method, but this is generally not done, probably because the method can be widely applied when conceived as a voucher but would be of very limited utility if enough were built into it to secure the POE. For a good example of the use of this method see Byrne [14].

Process tracing is, put simply, "the examination of intermediate steps in a process to make inferences about hypotheses on how that process took place and how it generated the outcome of interest." Bennett & Checkel [15, 6]. Done well (as Bennett and Checkel note) it does not just register intermediate steps but investigates *how* they came about, both the support factors necessary for each step to lead to the next and other factors that independently influence each step. If the investigation is expanded far enough, we end up with something that looks like a SCEM, which I discuss next. For a discussion of the methods and illustrations of their use , see Bennett & Checkel [15].

SCEM stands for "situation-specific causal equation model". These can ground evidence for a causal connection between a treatment and an outcome in the single case: a specific setting, population or individual. The SCEM includes the POE connecting treatment and outcome and also POEs for other causes of the outcome and further effects. It looks much like a causal structural model, except with a SCEM we do not look to statistics for help to fill values into the schema but investigate the actual case itself, as in a case study or in process tracing. As with casual Bayes nets, the idea is that by investigating a larger structure, we can sometimes get better evidence about the target POE, for instance by observing other effects that would follow had requisite intermediate steps occurred. Many types of evidence commonly used for singular casual connections, including most of what is mentioned by Bradford Hill [16], can be seen as testing aspects of a SCEM, including

- Cause characteristics
- Effect characteristics
- Symptoms of causality
- Presence of support factors
- Absence of derailers
- Presence of requisite intermediaries

Using the full SCEM allows us to see just why should count as *evidence* for the claim in view and just what role they play. For more on this, see Cartwright [17, 18].

5. What evidence should Guidelines consult?

The almost universal answer in EBM is "evidence from RCTs". What is their advantage? True, they are clinchers. But there are, as I have indicated, lots of other clinchers. And all clinchers are equal. If their assumptions are met, their conclusions follow deductively. Perhaps it is supposed that the assumptions for an RCT are generally more often met (or meetable) than those for other methods. What justifies that? Especially given that the easiest assumption to feel secure about for RCTs – that the assignment is done "randomly" – is far from enough to support orthogonality, which is itself only one among the assumptions that need support. I sometimes hear, "Only the RCT can control for unknown unknowns." But *nothing* can control for unknowns that we know nothing about. There is no reason to suppose that, for a given conclusion, the causal knowledge that it takes to stop post-randomization correlations in an RCT is always, or generally, more available or more reliable than the knowledge required for one or another of the other methods to be reliable.

It is also essential to be clear what the conclusion is. As with any study method, RCTS can only draw conclusions about the objects studied – for the RCT, the population enrolled in the trial, which is seldom the one we are interested in. The RCT method can be expanded of course to include among its assumptions that the trial population is a representative sample of the target. Then it follows deductively that the difference in mean outcomes between treatment and control groups is an unbiased estimate of the ATE of the target population. How often are we warranted in assuming that, though, and on what grounds? Without this assumption, an RCT is just a voucher for claims about any except the trial population. What then justifies placing it above methods that are clinchers for claims we are really interested in – about target populations?

Nor should we assume that clinchers are better than vouchers. A clincher *would be* an ideal source of evidence in any case where there were good reasons to suppose the relevant assumptions for its success were met. But if the assumptions aren't met, clinchers may provide no evidence at all. Agreed, the same is true of vouchers. But since vouchers aspire to less, often the assumptions they require are easier to meet.

6. Closing question: Why is the medical community so reluctant to admit other methods?

I have mentioned these other methods repeatedly at meetings and conferences and in conversation with EBM advocates. But I seldom note any uptake. It is not that the methods are given low grades as evidence for Guidelines and for practice. They do not even get into the discussion. Look for instance at the GRADE hierarchy [19], where none appear except correlational studies, which are treated as deficient wannabe RCTs. Or look at the Academy of Medical Sciences 2017 report, *Sources of evidence for assessing the safety, efficacy and effectiveness of medicines* [20]. I was one of a working group of 11 who met regularly to prepare that report. We did discuss a summary of various

of these methods reviewed here but none made it into the final document. Admittedly, each of these has its drawbacks and can only be used in cases where we have the requisite background knowledge. But that is true of RCTs as well. If we don't know how to police for the unknown unknowns post-random assignment, we can only view our results as correlational, not causal, and if we don't know the facts about causal structures that allow us to export from the trial to a target population, they are *irrelevant* to what we want to know. Other methods that deal with data from the target can often provide evidence for what we want to know about, and occasionally we are even in an epistemic position for these to clinch the conclusion. So ignoring evidence from these methods in constructing Guidelines seems daft. Worse, given what we need to know for individual care and how difficult it is to draw conclusions reliably even given a good amount of evidence , I think it is irresponsible.

References

1. Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*. In press. Available online: https://www.sciencedirect.com/science/article/pii/S0277953617307359.
2. Deaton, A, Cartwright N. Understanding and misunderstanding randomized controlled trials. *NBER Working Paper No. w22595 and Durham University: CHESS Working Paper No. 2016-05.* Available athttps://www.dur.ac.uk/resources/chess/CHESSK4UWP_2016_05_DeatonCartwright.pdf.
3. Pearl, J, & Bareinboim, E. External validity: From do-calculus to transportability across populations. *Statistical Science*. 2014; 29(4): 579-95.
4. Pearl J, Bareinboim E. Transportability of causal and statistical relations: A formal approach. In: *Proceedings of the 25th AAAI Conference on Artificial Intelligence.* Menlo Park, CA: AAAI Press; 2011: 247-254.
5. Newhouse JP, McClellan M. Econometrics in outcomes research: The Use of Instrumental Variables. *Annu. Rev. Public Health.* 1998; 19: 17–34.
6. Reiss, J. Causal Instrumental Variables and Interventions. *Philosophy of Science*. 2005: 72: 964–976.
7. Cartwright, N. Two Theorems on Invariance and Causality. *British Journal for the Philosophy of Science.* 2003; 70: 203-224.
8. Cartwright, N. *Hunting Causes and Using Them. Approaches in Philosophy and Economics.* Cambridge: Cambridge University Press; 2007.
9. Spirtes P. Introduction to Causal Inference. *Journal of Machine Learning Research.* 2010; 11: 1643-1662.
10. Shrier I, Platt R. Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology.* 2008; 8:70-79.
11. Aalen OO, Røysland K, Gran JM, Kouyos R, Lange T. Can we believe the DAGs? A comment on the relationship between causal DAGs and mechanisms. *Statistical Methods in Medical Research*. 2016; 25(5): 2294–2314.
12. Preuveneers D, Berbers Y. Mobile phones assisting with health self-care: a diabetes case study. In: *MobileHCI '08. Proceedings of the 10th international conference on Human computer interaction with mobile devices and services.* New York: ACM; 2008: 177-186. doi: 10.1145/1409240.1409260
13. Byrne D, Ragin, CC. *The SAGE Handbook of Case-Based Methods*. London: Sage; 2009.

14. Byrne D. Getting up - Staying Up? - Exploring Trajectories in Household Incomes Between 1992 and 2006. *Sociological Research Online.* 2012; 17(2). *<http://www.socresonline.org.uk/17/2/8.html> 10.5153/sro.2601.*
15. Bennett A, Checkel J. *Process Tracing. From metaphor to analytic tool.* Cambridge: Cambridge University Press; 2015.
16. Hill, A B. The environment and disease: Association or causation?. *Proceedings of the Royal Society of Medicine.* 1965; 58(5): 295-300.
17. Cartwright, N. Single Case Causes: What is Evidence and Why. In: Chao H, Chen S, Reiss J, eds. *Philosophy of Science in Practice: Nancy Cartwright and the Nature of Scientific Reasoning*. Dordrecht: Springer; 2016: 11-27. Also as *Durham University: CHESS Working Paper No. 2015-02.* 2015. Available at https://www.dur.ac.uk/resources/chess/CHESSWP_2015_02.pdf.
18. Cartwright, N. How to Learn about Causes in the Single Case. *Durham University: CHESS Working Paper No. 2017-04.* 2017. Available at https://www.dur.ac.uk/resources/chess/CHESSK4UWP_2017_04_Cartwright.pdf
19. Guyatt GH, Oxman AD, Kunz R, et al. What is "quality of evidence" and why is it important to clinicians? *BMJ*. 2008; 336: 995-998.
20. Rutter M, Breckenridge A, Cartwright N, et al. *Sources of evidence for assessing the safety, efficacy and effectiveness of medicines*. London: Academy of Medical Sciences; 2017.